

Tunable Ultrafast Thermal Relaxation in Graphene Measured by Continuous-Wave Photomixing

M. Mehdi Jadidi,^{1, a} Ryan J. Suess,¹ Cheng Tan,² Xinghan Cai,³ Kenji Watanabe,⁴ Takashi Taniguchi,⁴ Andrei B. Sushkov,⁵ Martin Mittendorff,¹ James Hone,² H. Dennis Drew,⁵ Michael S. Fuhrer,^{5, 6} and Thomas E. Murphy^{1, b}

¹*Institute for Research in Electronics & Applied Physics,
University of Maryland, College Park, MD 20742, USA*

²*Department of Mechanical Engineering, Columbia University, New York, New York 10027, USA*

³*Department of Physics, University of Washington, Seattle, Washington 98195, USA*

⁴*National Institute for Materials Science, 1-1 Namiki, Tsukuba 305-0044, Japan*

⁵*Center for Nanophysics and Advanced Materials,
University of Maryland, College Park, Maryland 20742, USA*

⁶*School of Physics and Astronomy, Monash University, 3800 Victoria, Australia*

Hot electron effects in graphene are significant because of graphene's small electronic heat capacity and weak electron-phonon coupling, yet the dynamics and cooling mechanisms of hot electrons in graphene are not completely understood. We describe a novel photocurrent spectroscopy method that uses the mixing of continuous-wave lasers in a graphene photothermal detector to measure the frequency dependence and nonlinearity of hot-electron cooling in graphene as a function of the carrier concentration and temperature. The method offers unparalleled sensitivity to the nonlinearity, and probes the ultrafast cooling of hot carriers with an optical fluence that is orders of magnitude smaller than in conventional time-domain methods, allowing for accurate characterization of electron-phonon cooling near charge neutrality. Our measurements reveal that near the charge neutral-point the nonlinear power dependence of the electron cooling is dominated by disorder-assisted collisions, while at higher carrier concentrations conventional momentum-conserving cooling prevails in the nonlinear dependence. The relative contribution of these competing mechanisms can be electrostatically tuned through the application of a gate voltage – an effect that is unique to graphene.

When graphene absorbs electromagnetic radiation, its electrons heat up and produce a measurable thermoelectric response, even at room temperature. Because of graphene's gapless dispersion relation, small electronic heat capacity, and anomalously weak electron-phonon coupling, this photothermal detection mechanism is broadband (from dc to visible), highly sensitive, and fast [1–5]. The speed, temperature dependence, and power dependence of these detectors depend critically upon how fast and by what mechanisms the hot carriers relax [6–8]. Two primary cooling mechanisms have been identified: supercollision cooling, in which disorder-assisted scattering allows for non-momentum-conserving transitions, and conventional momentum-conserving electron-phonon cooling [6, 7, 9–14]. Evidence for conventional momentum-conserving cooling (which is linear with temperature) has been observed only at low temperatures in high-quality graphene [8]. In experimental measurements, the cooling process is inferred from how the photothermal response depends on temperature, power, or time for either pulsed or continuous-wave illumination. Time-domain methods that are used to study thermal relaxation dynamics typically employ intense optical pulses, which significantly disturb the electron temperature, and can in some cases excite higher energy optical phonons in addition to acoustic phonons [7, 15–17]. Moreover, as we show here, the factors that govern the power depen-

dence of the photothermal response can be different from those that determine the cooling rate. It has been shown that, uniquely in graphene, the relative strength of the two competing cooling channels can be controlled by the carrier concentration [6, 8, 12].

Here we employ a new nonlinear photomixing method to simultaneously quantify the nonlinearity in the photoreponse and the carrier-density dependence of electron cooling in graphene. This method easily distinguishes between sublinear and superlinear power dependence, which indicate supercollision cooling and conventional cooling, respectively. Our measurements show that while supercollision cooling dominates the nonlinear response near the charge neutral point, at higher carrier densities, conventional cooling is the dominant contribution to the nonlinearity. Furthermore, we show that when two detuned near-IR lasers co-illuminate the graphene, the resulting dc photovoltage depends upon their heterodyne difference frequency. This enables the direct measurement of the electron cooling rate in the frequency domain with orders of magnitude weaker optical excitation (smaller temperature rise) than traditional time-domain methods, by simply tuning the wavelength of one of the continuous-wave lasers.

Figure 1(a) depicts the heterodyne photomixing setup used here to characterize the photothermal response of graphene. Two fiber-coupled continuous-wave near-IR lasers, one wavelength tunable ($\lambda_1 = 1540\text{--}1565\text{ nm}$) and one at fixed-wavelength ($\lambda_2 = 1545\text{ nm}$), were amplified, spatially combined, polarized, and focused using an aspheric lens to a $3\text{ }\mu\text{m}$ spot on the graphene channel. The

^a mmjadidi@umd.edu

^b tem@umd.edu

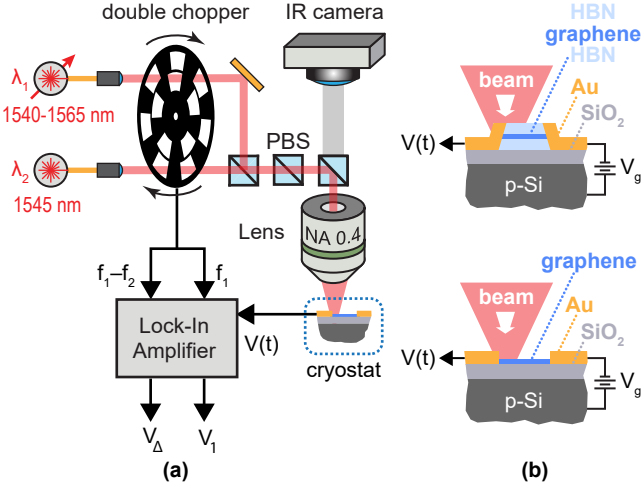


FIG. 1. Diagram of heterodyne photomixing experiment. (a) Two near-IR continuous-wave beams, one with a tunable wavelength, are modulated at two different frequencies f_1 and f_2 , then overlapped and focused down to the graphene photodetector. The photovoltages produced by laser 1 (V_1) and the mixing of two beams (V_Δ) are detected at the modulation frequencies f_1 and $f_1 - f_2$ respectively. (b) Diagram of the HBN-encapsulated graphene (top) and exfoliated graphene (bottom) photodetector devices. The optical beams are illuminated close to one of the metal contacts, and the carrier density of graphene is altered by applying an electrostatic voltage (V_g) between the doped silicon substrate and graphene.

position of the focused beam was chosen to maximize the photovoltage, which occurs when the beam is focused close to one of the contacts [18, 19]. The combined optical power illuminating the first (second) device was about 6 (2.1) mW, from which we estimate the total absorbed intensity to be $I = 850$ (300) W/cm². The graphene photodetector device was held in a liquid helium cryostat with short working distance optical access to controllably vary the lattice temperature T_L between 10 K and room temperature. The two lasers were mechanically chopped using a twin-slot (5/7) chopping wheel at frequencies $f_1 = 500$ Hz and $f_2 = 700$ Hz. The photovoltage was synchronously detected at both f_1 and the difference $f_2 - f_1$, using a dual-reference digital lock-in amplifier (Signal Recovery 7270), which simultaneously records the photovoltages V_1 and V_Δ . The phases of the two lock-in detection channels were calibrated to produce the correct sign, relative to one another. Measurements were performed as a function of the gate voltage V_g , and the optical difference frequency $\Delta\nu = \Omega/2\pi$, which was swept from -0.6 to $+2.5$ THz by tuning laser 1.

To better elucidate the role of disorder, we considered two different graphene detectors shown in Fig. 1(b): one using an edge-contacted hexagonal boron nitride (HBN) encapsulated graphene channel[20, 21], and a second fabricated from an unencapsulated exfoliated flake[5]. The Supplemental Material[33] details the fabrication and dc

electrical characterization of the devices (Sec. S4).

The electron temperature T in the graphene evolves according to the nonlinear differential equation [6, 9, 22]

$$\alpha T \frac{dT}{dt} + \beta_1(T - T_L) + \beta_3(T^3 - T_L^3) = I(t) \quad (1)$$

where T_L is the lattice temperature, αT is the specific heat of the graphene carriers, the coefficients β_1 and β_3 are the rate coefficients for the conventional and super-collision cooling mechanisms, respectively, and $I(t)$ is the absorbed near-infrared optical intensity. For the two-laser illumination shown in Fig. 1(a), the absorbed intensity is $I(t) = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \Omega t$, where I_1 and I_2 represent the absorbed intensities of lasers 1 and 2, respectively and $\Omega \equiv 2\pi c(\lambda_2^{-1} - \lambda_1^{-1})$ is their heterodyne difference frequency.

It is assumed that the electrons are in the degenerate regime ($E_F \gg k_B T$), and that the electron-electron collisions are fast enough ($\tau_{ee}^{-1} \gg \Omega$) to allow the temperature of the electron gas to be well defined [9, 23]. The hot-electron diffusion length is $\xi = (\kappa/\gamma\alpha T)^{1/2} = v_F(\gamma\Gamma)^{-1/2}$, where κ is the electronic thermal conductivity, Γ is the carrier scattering rate, and the Wiedemann-Franz law was used in the second equality. Even for the encapsulated device considered here, by estimating Γ from dc measurements in Fig. S3, we estimate that $500 \text{ nm} < \xi < 1.5 \text{ }\mu\text{m}$, which is smaller than the optical beam size employed. We therefore ignore spatial inhomogeneity in $I(t)$ and the thermal diffusion of hot carriers out of the laser beam.

The three model parameters α , β_1 , and β_3 appearing in Eq. (1) depend implicitly on the Fermi level E_F (determined by gating) and the disorder mean-free path l (related to the quality of the graphene) as $\alpha = 2\pi k_B^2 E_F / (3\hbar^2 v_F^2)$, $\beta_1 = V_D^2 E_F^4 k_B / (2\pi \rho \hbar^5 v_F^6)$, $\beta_3 = \zeta(3) V_D^2 E_F k_B^3 / (\pi^2 \rho \hbar^4 v_F^3 s^2 l)$, where v_F is the Fermi velocity, ρ is the areal mass density, s is the speed of sound in graphene, $\zeta(3) \simeq 1.202$ is the Riemann zeta function, and V_D is the acoustic deformation potential. We note that the substrate surface polar phonons may also play a role in hot electron cooling in graphene[24–26], and their effect on the photoresponse can be regarded as a linear cooling term (β_1) in Eq. (1)[27]. At temperatures far below the Bloch-Grüneisen temperature Eq. (1) must be modified to include a cooling term proportional to T^4 [22]. We estimate that even at the lowest temperatures and carrier concentrations experimentally considered here ($T = 25$ K, $E_F \sim 60$ meV), the measurement temperature matches or exceeds the Bloch-Grüneisen temperature.

The resulting photothermoelectric voltage V produced by the Seebeck effect is then related to the electron temperature by $V = rT(T - T_L)$, where rT is the Seebeck coefficient of graphene[2, 28]. This nonlinear relationship between temperature and photovoltage could be generalized to include a nonlinearity in the Seebeck coefficient[29, 30], but the temperature dependence and power dependence of the observed nonlinearity indicate

that this effect is small in comparison to the nonlinearity in cooling. Although other photoresponse mechanisms, such as the photoelectric effect[31], might also contribute to the graphene photoresponse, a photothermoelectric model can adequately describe the photoresponse at the graphene-metal interface[5, 13, 32].

Equation (1) can be solved using a power series expansion (Supplemental Material[33], Sec. S1), and the resulting dc photovoltage is found to be

$$\begin{aligned} V(I_1, I_2) = & a_1(I_1 + I_2) + a_2(I_1^2 + I_2^2) - a_3(I_1^3 + I_2^3) \dots \\ & + 2a_2I_1I_2\left(1 + \frac{\gamma^2}{\Omega^2 + \gamma^2}\right) \dots \\ & - 3a_3I_1I_2(I_1 + I_2)\left(1 + \frac{2\gamma^2}{\Omega^2 + \gamma^2}\right) \end{aligned} \quad (2)$$

where $\gamma \equiv (\beta_1 + 3\beta_3T_L^2)/\alpha T_L$ is the linearized cooling rate from both mechanisms. The coefficients a_1 , a_2 , and a_3 are given by

$$a_1 \equiv \frac{r}{\alpha\gamma}, \quad a_2 \equiv \frac{r\beta_1}{(\alpha\gamma T_L)^3}, \quad a_3 \equiv \frac{3r\beta_3^2}{T_L^2\alpha^5\gamma^5} \quad (3)$$

The final two terms in Eq. (2) which contain the factor I_1I_2 , represent a nonlinear interaction of the two beams, which occurs only when both beams are present. In order to sensitively detect only these mixing products, we employ a double-modulation configuration in which laser 1 is mechanically chopped at a frequency f_1 , laser 2 is chopped at f_2 , and the photovoltage V is synchronously detected using a lock-in amplifier at the chopping difference frequency $\Delta f \equiv f_1 - f_2$ (not to be confused with the heterodyne difference frequency). The resulting photovoltage $V_\Delta \equiv V(I_1, I_2) - V(I_1, 0) - V(0, I_2)$ can be positive or negative, depending on the nonlinearity in the photothermal response. We also simultaneously measure the Fourier component at f_1 , denoted $V_1 \equiv V(I_1, 0)$, which gives the photovoltage produced by laser 1 alone.

By simply comparing the magnitude of the two terms that compose the linearized cooling rate γ , one can determine a condition for which process makes the largest contribution to the cooling rate. For nearly all of the experimental cases considered here and reported elsewhere, the cooling rate is largely limited by the supercollision term. Equation (3) reveals that despite this, the photoresponse can be either superlinear or sublinear in intensity, depending on the carrier density and graphene quality. As explained below, neither cooling effect can be ignored when analyzing the nonlinearity of the response.

When the heterodyne frequency exceeds the cooling rate ($\Omega \gg \gamma$), Eq. (2) simplifies to $V(I) = a_1I + a_2I^2 - a_3I^3$, where $I \equiv I_1 + I_2$ is the total absorbed optical intensity. The quadratic and cubic terms have opposite sign, and therefore describe superlinear or sublinear dependence on the optical intensity. From Eq. (3), one sees that the superlinear coefficient is proportional to β_1 , which we associate with momentum-conserving cooling,

while the sublinear coefficient is proportional to β_3 , which arises from supercollision cooling.

Figure 2(a) plots V_1 (black) and V_Δ (green) as a function of the gate voltage for the HBN-encapsulated device. These measurements were performed with $\Omega/2\pi = 2.5$ THz, which is much faster than the expected cooling rate at room temperature. The sign of the photothermal voltage V_1 depends on the gate voltage, as expected from the photothermoelectric effect [2, 3, 5]. For carrier densities near the charge neutral point, V_1 and V_Δ have opposite sign (as indicated by the blue shading), revealing a sub-linear power dependence, characteristic of supercollision cooling [7, 13]. In this regime the Fermi surface is small, and the allowed phonon energy space for the momentum-conserving collision is strongly constrained, thereby suppressing conventional electron-phonon cooling [6, 12]. At higher carrier densities, the behavior changes to super-linear (red shading), indicating that conventional cooling becomes stronger and dominates the photothermal nonlinearity. Figure 2(b) plots the single-beam photovoltage as a function of the incident optical power, confirming the sub- and superlinear behavior, respectively. Figure 2(c) illustrates the two cooling mechanisms schematically in k space, along with the predicted sublinear and superlinear power dependence. The transitions outside of the Dirac cone represent supercollision cooling, in which the spatial disorder in the graphene compensates for the electron-phonon momentum mismatch.

The threshold between these two nonlinear regimes can be approximated by equating the opposing terms in V_Δ , which gives

$$2a_2 \gtrless 3a_3I \quad (4)$$

where the upper and lower inequalities describe the conditions under which conventional cooling or supercollision cooling prevails in the nonlinear response, respectively. The relative importance of the two competing cooling channels depends on temperature, intensity, the carrier concentration (Fermi level), and indirectly on the material quality, which is related to the disorder mean-free path l . Even though the linearized cooling rate γ is limited by supercollision cooling, both effects are evident in the nonlinear response reported here.

In the Supplemental Material (Ref. [33], Sec. S2), we present the results of a similar measurement performed on lower-mobility exfoliated graphene on SiO_2 . Similar to the HBN-encapsulated device, we observe an expected transition from supercollision cooling to conventional cooling. The transition happens around $E_F = 80$ meV, and we use Eq. (4) to determine the ratio of the two rate coefficients, $\beta_1/\beta_3 = 5300 \text{ K}^2$. At room temperature, the supercollision contribution to the cooling rate γ is nearly $50\times$ larger than the contribution from conventional cooling. Despite this, both effects have a non-negligible role in the nonlinearity of the photoresponse, and their relative significance depends on the carrier density.

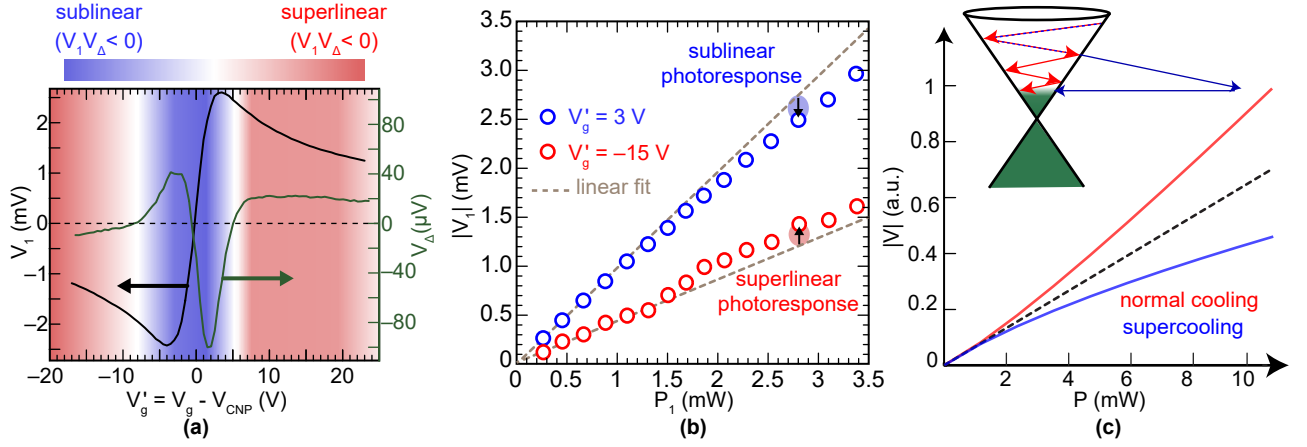


FIG. 2. Photovoltage versus carrier concentration and incident power, measured at room temperature. (a) Single-laser photovoltage (V_1), and nonlinear photomixing signal (V_Δ) measured versus the applied gate voltage V_g . V_{CNP} denotes the charge neutral point. The red ($V_1 V_\Delta > 0$) and blue ($V_1 V_\Delta < 0$) regions indicate the conditions where super- and sublinear power dependence is observed. (b) Measured photovoltage V_1 of a single laser versus incident optical power at two different gate voltages, showing sublinear and superlinear behavior. (c) Calculated photothermoelectric voltage (in arbitrary units) versus input optical power for the case of $\beta_1 = 0$ (blue) and $\beta_3 = 0$ (red), illustrating the sublinear and superlinear behavior, respectively. Inset: energy diagram illustrating the two different cooling mechanisms.

When the heterodyne difference frequency Ω is comparable to or smaller than the cooling rate γ , the electron temperature can follow the interferometric beating of the two lasers, which produces a larger photothermal voltage than when the lasers are widely detuned. The final two terms in Eq. (2) reveal that the nonlinear mixing signal V_Δ exhibits a Lorentzian dependence on the heterodyne difference frequency Ω , with a spectral width that is proportional to the cooling rate γ . As before, the double-chopping configuration allows for sensitive detection of this heterodyne photomixing signal.

Figure 3(a) plots the measured photovoltage V_Δ as a function of the gate voltage and heterodyne difference frequency, for the nonencapsulated graphene detector. In addition to the expected gate-voltage dependence discussed previously, the photoresponse exhibits a distinct spectral peak around $\Omega = 0$. Figure 3(b) shows the photomixing spectrum at a fixed gate voltage, along with the best-fit Lorentzian curve. From the linewidth, we estimate a cooling time of $\gamma^{-1} = 1.42$ ps, which is consistent with the time-domain pulse coincidence measurements [5, 37] reported for similar devices.

In order to confirm the thermal model for the photomixing, we repeated the heterodyne spectral measurements at temperatures from room temperature down to 25 K for the exfoliated sample on SiO_2 near the charge neutral point. As shown in Fig. 4(a), in all cases the photomixing signal exhibits a Lorentzian spectral dependence, but with a spectral width that decreases with temperature, as expected. The solid blue curve in Fig. 4(b) shows a fit to the linearized cooling rate, based on the model presented here. For the data points above $T_L = 80\text{ K}$, and for the conditions at the charge neutral point, the assumption of $E_F \gg k_B T$ (degenerate regime) is no

longer valid, which requires a modification of the cooling rate (Supplemental Material[33], Sec. S3). We therefore excluded these points when fitting the blue curve. However, when the parameters from the low-temperature fit were incorporated into the modified thermal model, it correctly predicts the observed high-temperature asymptotic behavior, indicated by the red curve, with no additional free parameters.

This nonlinear heterodyne photovoltage spectroscopy method has two important advantages over the traditional time-domain measurement using pulse coincidence [5, 7, 37]: (i) the frequency range and resolution is limited only by the tuning range and resolution of the laser, while in time-domain measurements the response is limited by the optical pulse width and repetition period; (ii) continuous-wave illumination produces a far smaller thermal stimulus to the graphene electrons than intense ultrafast pulses, thereby allowing the measurement of the temporal dynamics and nonlinearity of photodetection under low photothermal excitation for which the electron temperature is near the lattice temperature.

The model and measurements described here show that there are two competing cooling channels for hot electrons in graphene, and Eq. (4) describes the relative importance of each in the nonlinear response. In time-domain experiments reported elsewhere, the instantaneous absorbed intensity is orders of magnitude higher than the continuous-wave illumination considered here, in which case Eq. (4) predicts that supercollision cooling is the dominant contribution to the nonlinearity at all practically attainable doping concentrations. Moreover nonencapsulated graphene samples have a much smaller disorder mean-free path l , which further contributes to the relative importance of supercollision cooling over con-

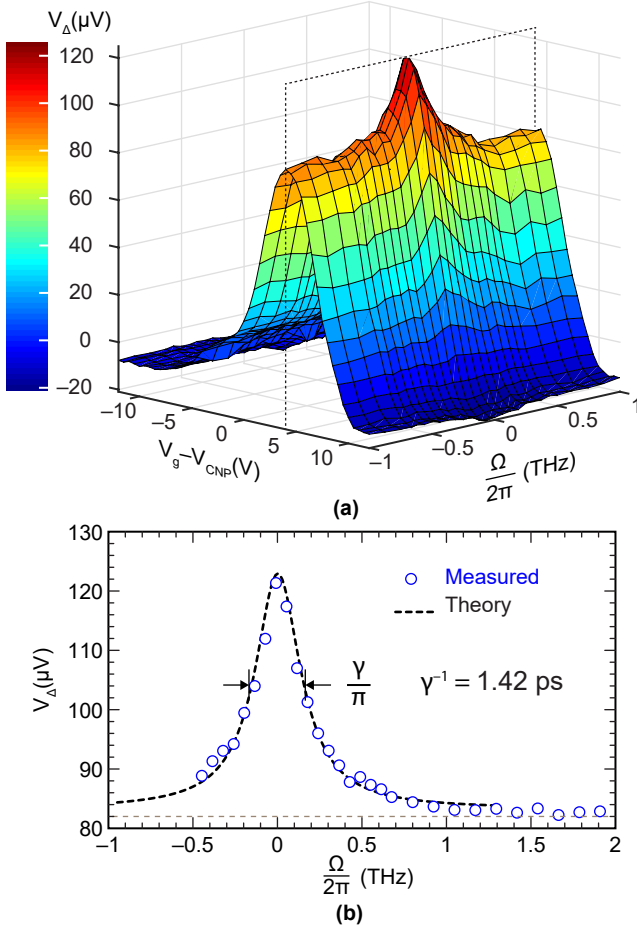


FIG. 3. Heterodyne photomixing response for the nonencapsulated device, measured at room temperature. (a) The nonlinear photomixing signal (V_{Δ}) as a function of the difference frequency (Ω) and gate voltage (V_g). At each gate voltage, V_{Δ} exhibits a Lorentzian-shaped dependence on the heterodyne difference frequency Ω . (b) V_{Δ} vs Ω for $V_g = 4$ V. The dashed curve is the theoretical calculation of the photomixing voltage based on a photothermoelectric effect. From the Lorentzian fit, the hot electron cooling time is estimated to be $\gamma^{-1} = 1.42$ ps.

ventional momentum-conserving cooling. In these cases, the photothermal response is often adequately described by supercollision cooling alone, for a wide range of carrier densities and temperatures [13, 38, 39]. For continuous-wave measurements on encapsulated devices, Eq. (4) also predicts that at sufficiently low temperatures, conventional cooling will prevail, consistent with temperature-dependent measurements reported recently[8].

We show that nonlinearity in the photothermoelectric effect causes photomixing when graphene is illuminated by near-infrared beams, and we describe a new heterodyne spectroscopy method that accurately measures this nonlinearity in the frequency domain. Exceedingly small nonlinearities in the photoresponse can be probed using continuous-wave illumination, which accurately elucidates the physical mechanisms behind the nonlinear-

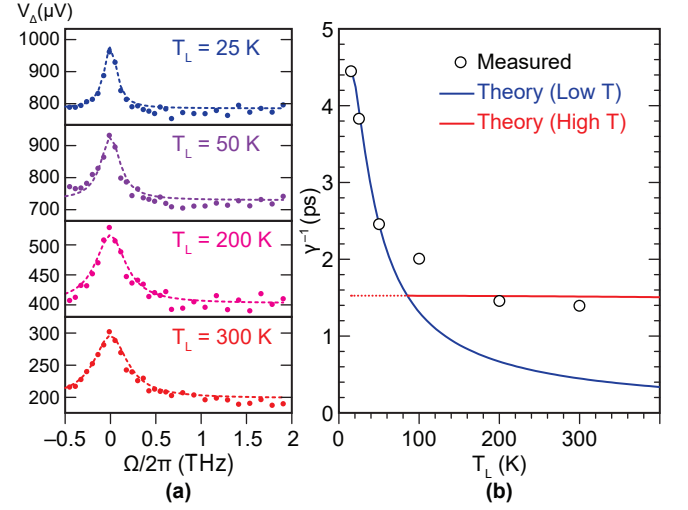


FIG. 4. Temperature dependence of the cooling rate. (a) The two-beam mixing signal as a function of the difference frequency measured close to the charge neutral point (sublinear power dependence regime) at different temperatures for the exfoliated sample on SiO_2 . The dashed curves are Lorentzian fits. (b) The black circles are the extracted hot electron cooling time constant $\tau \equiv \gamma^{-1}$ from the Lorentzian fits in panel (a). The blue (red) curve is the theory fit for low (high) temperature.

ity and cooling. In particular, the measurements reveal the role that disorder plays in the cooling of hot electrons, and the interplay of different cooling channels at different carrier concentrations. The method permits the direct measurement of the cooling rate in graphene using swept laser spectroscopy, which offers several advantages over pump-probe or pulse-coincidence measurements. The work also implies that nonlinear photomixing in graphene is very promising for the development of new optical and THz photomixing devices.

This work was sponsored by the U.S. ONR (N000141310865) and the U.S. NSF (ECCS 1309750). C.T. was supported by a DOD-AFOSR, NDSEG fellowship under Contract No. FA9550-11-C-0028, 32 CFR 168a. C.T. and J.H. acknowledge financial support from the Nanoelectronics Research Initiative (NRI) through the Institute for Nanoelectronics Discovery and Exploration (INDEX). K.W. and T.T. acknowledge support from the Elemental Strategy Initiative conducted by the MEXT, Japan and a Grant-in-Aid for Scientific Research on Innovative Areas “Science of Atomic Layers” from JSPS. M.S.F. was supported in part by an ARC Laureate Fellowship.

REFERENCES

- [1] Y.M. Zuev, W. Chang, and P. Kim, “Thermoelectric and magnetothermoelectric transport measurements of graphene.” *Phys. Rev. Lett.* **102**, 096807 (2009).
- [2] X. Xu, N.M. Gabor, J.S. Alden, A.M. van der Zande,

- and P.L. McEuen, “Photo-Thermoelectric Effect at a Graphene Interface Junction,” *Nano Lett.* **10**, 562 (2010).
- [3] N.M. Gabor, J.C. Song, Q. Ma, N.L. Nair, T. Taychatanapat, K. Watanabe, T. Taniguchi, L.S. Levitov, and P. Jarillo-Herrero, “Hot carrier-assisted intrinsic photoresponse in graphene,” *Science* **334**, 648 (2011).
 - [4] F.H.L. Koppens, T. Mueller, P. Avouris, A.C. Ferrari, M.S. Vitiello, and M. Polini, “Photodetectors based on graphene, other two-dimensional materials and hybrid systems,” *Nat. Nanotechnol.* **9**, 780 (2014).
 - [5] X. Cai, A. B. Sushkov, R. J. Suess, M. M. Jadidi, G. S. Jenkins, L. O. Nyakiti, R. L. Myers-Ward, S. Li, J. Yan, D. K. Gaskill, T. E. Murphy, H. D. Drew, and M. S. Fuhrer, “Sensitive room-temperature terahertz detection via the photothermoelectric effect in graphene,” *Nat. Nanotechnol.* **9**, 814 (2014).
 - [6] J. C. W. Song, M. Y. Reizer, and L. S. Levitov, “Disorder-Assisted Electron-Phonon Scattering and Cooling Pathways in Graphene,” *Phys. Rev. Lett.* **109**, 106602 (2012).
 - [7] M.W. Graham, S.F. Shi, D.C. Ralph, J. Park, and P.L. McEuen, “Photocurrent measurements of supercollision cooling in graphene,” *Nat. Phys.* **9**, 103 (2013).
 - [8] Q. Ma, N.M. Gabor, T.I. Andersen, N.L. Nair, K. Watanabe, T. Taniguchi, and P. Jarillo-Herrero, “Competing channels for hot-electron cooling in graphene,” *Phys. Rev. Lett.* **112**, 247401 (2014).
 - [9] R. Bistritzer and A. H. MacDonald, “Electronic Cooling in Graphene,” *Phys. Rev. Lett.* **102**, 206410 (2009).
 - [10] W.-K. Tse and S. Das Sarma, “Energy relaxation of hot Dirac fermions in graphene,” *Phys. Rev. B* **79**, 235406 (2009).
 - [11] A. C. Betz, F. Vialla, D. Brunel, C. Voisin, M. Picher, A. Cavanna, A. Madouri, G. Fève, J.-M. Berroir, B. Plaçais, and E. Pallecchi, “Hot Electron Cooling by Acoustic Phonons in Graphene,” *Phys. Rev. Lett.* **109**, 056805 (2012).
 - [12] A. C. Betz, F. Vialla, D. Brunel, C. Voisin, M. Picher, A. Cavanna, A. Madouri, G. Fève, J.-M. Berroir, B. Plaçais, and E. Pallecchi, “Supercollision cooling in undoped graphene,” *Nat. Phys.* **9**, 109 (2013).
 - [13] R.J. Shiue, Y. Gao, Y. Wang, C. Peng, A.D. Robertson, D.K. Efetov, S. Assefa, F.H. Koppens, J. Hone, and D. Englund, “High-Responsivity Graphene/Boron Nitride Photodetector and Autocorrelator in a Silicon Photonic Integrated Circuit,” *Nano Lett.* **15**, 7288 (2015).
 - [14] C. B. McKitterick, D. E. Prober, and M. J. Rooks, “Electron-phonon cooling in large monolayer graphene devices,” *Phys. Rev. B* **93**, 075410 (2016).
 - [15] H. Wang, J. H. Strait, P. A. George, S. Shivaraman, V. B. Shields, M. Chandrashekhar, J. Hwang, F. Rana, Spencer M. G., C. S. Ruiz-Vargas, and J. Park, “Ultrafast relaxation dynamics of hot optical phonons in graphene,” *Appl. Phys. Lett.* **96**, 081917 (2010).
 - [16] S. Winnerl, M. Orlita, P. Plochocka, P. Kossacki, M. Potemski, T. Winzer, E. Malic, A. Knorr, M. Sprinkle, C. Berger, W. A. de Heer, H. Schneider, and M. Helm, “Carrier Relaxation in Epitaxial Graphene Photoexcited Near the Dirac Point,” *Phys. Rev. Lett.* **107**, 237401 (2011).
 - [17] K. J. Tielrooij, J. C. W. Song, S. A. Jensen, A. Centeno, A. Pesquera, A. Zurutuza Elorza, M. Bonn, L. S. Levitov, and F. H. L. Koppens, “Photoexcitation cascade and multiple hot-carrier generation in graphene,” *Nat. Phys.* **9**, 248 (2013).
 - [18] E.J. Lee, K. Balasubramanian, R.T. Weitz, M. Burghard, and K. Kern, “Contact and edge effects in graphene devices,” *Nat. Nanotechnol.* **3**, 486 (2008).
 - [19] F. Xia, T. Mueller, R. Golizadeh-Mojarad, M. Freitag, Y.M. Lin, J. Tsang, V. Perebeinos, and P. Avouris, “Photocurrent imaging and efficient photon detection in a graphene transistor,” *Nano Lett.* **9**, 1039 (2009).
 - [20] C. R. Dean, A. F. Young, I. Meric, C. Lee, L. Wang, S. Sorgenfrei, K. Watanabe, T. Taniguchi, P. Kim, K. L. Shepard, and J. Hone, “Boron nitride substrates for high-quality graphene electronics,” *Nat. Nanotechnol.* **5**, 722 (2010).
 - [21] L. Wang, I. Meric, P. Y. Huang, Q. Gao, Y. Gao, H. Tran, T. Taniguchi, K. Watanabe, L. M. Campos, D. A. Muller, J. Guo, P. Kim, J. Hone, K. L. Shepard, and C. R. Dean, “One-Dimensional Electrical Contact to a Two-Dimensional Material,” *Science* **342**, 614 (2013).
 - [22] J. K. Viljas and T. T. Heikkilä, “Electron-phonon heat transfer in monolayer and bilayer graphene,” *Phys. Rev. B* **81**, 245404 (2010).
 - [23] S. Das Sarma, S. Adam, E. H. Hwang, and E. Rossi, “Electronic transport in two-dimensional graphene,” *Rev. Mod. Phys.* **83**, 407 (2011).
 - [24] T. Low, V. Perebeinos, R. Kim, M. Freitag, and P. Avouris, “Cooling of photoexcited carriers in graphene by internal and substrate phonons,” *Phys. Rev. B* **86**, 045413 (2012).
 - [25] E. H. Hwang and S. Das Sarma, “Surface polar optical phonon interaction induced many-body effects and hot-electron relaxation in graphene,” *Phys. Rev. B* **87**, 115432 (2013).
 - [26] A. Principi, M. B. Lundeberg, N.C.H. Hesp, K.-J. Tielrooij, F.H.L. Koppens, and M. Polini, “Super-Planckian electron cooling in a van der Waals stack,” *arXiv:1608.01516*.
 - [27] M. Freitag, T. Low, and P. Avouris, “Increased Responsivity of Suspended Graphene Photodetectors,” *Nano Lett.* **13**, 1644–1648 (2013).
 - [28] J.C. Song, M.S. Rudner, C.M. Marcus, and L.S. Levitov, “Hot carrier transport and photocurrent response in graphene,” *Nano Lett.* **11**, 4688–4692 (2011).
 - [29] J. Duan, X. Wang, X. Lai, G. Li, K. Watanabe, T. Taniguchi, M. Zebbarjadi, and E. Y. Andrei, “High Thermoelectric Power Factor in Graphene/hBN Devices,” (2016), *arXiv:1607.00583 [cond-mat.mes-hall]*.
 - [30] E. H. Hwang, E. Rossi, and S. Das Sarma, “Theory of thermopower in two-dimensional graphene,” *Phys. Rev. B* **80**, 235415 (2009).
 - [31] T. J. Echtermeyer, P. S. Nene, M. Trushin, R. V. Gorbachev, A. L. Eiden, S. Milana, Z. Sun, J. Schliekmann, E. Lidorikis, K. S. Novoselov, and A. C. Ferrari, “Photothermoelectric and Photoelectric Contributions to Light Detection in Metal/Graphene/Metal Photodetectors,” *Nano Lett.* **14**, 3733 (2014).
 - [32] K. J. Tielrooij, M. Massicotte, L. Piatkowski, A. Woessner, Q. Ma, P. Jarillo-Herrero, N.F. van Hulst, and F.H.L. Koppens, “Hot-carrier photocurrent effects at graphenemetal interfaces,” *J. Phys. Condens. Matter* **27**, 164207 (2015).
 - [33] See Supplemental Material, which cites Refs. [34–36], for details of thermal model, device fabrication, and electrical characterization.
 - [34] P.A. Khomyakov, A.A. Starikov, G. Brocks, and

- P.J. Kelly, “Nonlinear screening of charges induced in graphene by metal contacts,” *Phys. Rev. B* **82**, 115437 (2010).
- [35] T. Mueller, F. Xia, M. Freitag, J. Tsang, and P. Avouris, “Role of contacts in graphene transistors: A scanning photocurrent study,” *Phys. Rev. B* **79**, 245430 (2009).
- [36] C. H. Lui, K. F. Mak, J. Shan, and T. F. Heinz, “Ultrafast Photoluminescence from Graphene,” *Phys. Rev. Lett.* **105**, 127404 (2010).
- [37] D. Sun, G. Aivazian, A.M. Jones, J.S. Ross, W. Yao, D. Cobden, and X. Xu, “Ultrafast hot-carrier-dominated photocurrent in graphene,” *Nat. Nanotechnol.* **7**, 114 (2012).
- [38] M.W. Graham, S.F. Shi, Z. Wang, D.C. Ralph, J. Park, and P.L. McEuen, “Transient absorption and photocurrent microscopy show that hot electron supercollisions describe the rate-limiting relaxation step in graphene,” *Nano Lett.* **13**, 5497 (2013).
- [39] M. M. Jadidi, J. C. Konig-Otto, A. B. Sushkov, S. Winnerl, H. D. Drew, T. E. Murphy, and M. Mitterdorff, “Nonlinear Terahertz Absorption of Graphene Plasmons,” *Nano Lett.* **16** (2016).

Supplemental Material

S1. NONLINEAR THERMAL MODEL

The electron temperature in the graphene may be modeled by the following nonlinear differential equation:

$$\alpha T \frac{dT}{dt} + \beta_1(T - T_L) + \beta_3(T^3 - T_L^3) = I(t) \quad (S1)$$

where T represents the graphene electron temperature, T_L is the lattice temperature, and $I(t)$ is the absorbed optical power per unit area. αT is the specific heat in the graphene and the terms proportional to β_1 and β_3 describe momentum-conserving cooling and disorder-assisted supercollision cooling, respectively.

We re-write these equations in terms of $x \equiv T - T_L$, the deviation from the lattice temperature:

$$\alpha(T_L + x) \frac{dx}{dt} + \beta_1 x + \beta_3 [(T_L + x)^3 - T_L^3] = I(t) \quad (S2)$$

We next assume that $x \ll T_L$, i.e., the photoinduced change in electron temperature is small in comparison to the equilibrium (lattice) temperature. With this assumption, $x(t)$ may be expanded in a power series in the intensity I ,

$$x(t) = x^{(1)}(t) + x^{(2)}(t) + x^{(3)}(t) + \dots \quad (S3)$$

Where $x^{(n)} \propto I^n$, and we are retaining terms up to third order. Substituting this expansion into Eq. (S2) gives

$$\alpha(T_L + x^{(1)} + x^{(2)} + x^{(3)}) \frac{d}{dt}(x^{(1)} + x^{(2)} + x^{(3)}) + \beta_1(x^{(1)} + x^{(2)} + x^{(3)}) + \beta_3 [(T_L + x^{(1)} + x^{(2)} + x^{(3)})^3 - T_L^3] = I(t) \quad (S4)$$

Next, we expand Eq. (S4) and separately equate the orders to obtain the following inhomogeneous linear differential equations for $x^{(1)}$, $x^{(2)}$ and $x^{(3)}$,

$$\alpha T_L \frac{dx^{(1)}}{dt} + (\beta_1 + 3\beta_3 T_L^2)x^{(1)} = I(t) \quad (S5)$$

$$\alpha T_L \frac{dx^{(2)}}{dt} + (\beta_1 + 3\beta_3 T_L^2)x^{(2)} = -\alpha x^{(1)} \frac{dx^{(1)}}{dt} - 3\beta_3 T_L [x^{(1)}]^2 \quad (S6)$$

$$\alpha T_L \frac{dx^{(3)}}{dt} + (\beta_1 + 3\beta_3 T_L^2)x^{(3)} = -\alpha x^{(1)} \frac{dx^{(2)}}{dt} - \alpha x^{(2)} \frac{dx^{(1)}}{dt} - 6\beta_3 T_L x^{(1)} x^{(2)} - \beta_3 [x^{(1)}]^3 \quad (S7)$$

which can be re-written as:

$$\frac{dx^{(1)}}{dt} + \gamma x^{(1)} = \frac{I(t)}{\alpha T_L} \quad (S8)$$

$$\frac{dx^{(2)}}{dt} + \gamma x^{(2)} = -\frac{1}{T_L} x^{(1)} \frac{dx^{(1)}}{dt} - \frac{3\beta_3}{\alpha} [x^{(1)}]^2 \quad (S9)$$

$$\frac{dx^{(3)}}{dt} + \gamma x^{(3)} = -\frac{1}{T_L} x^{(1)} \frac{dx^{(2)}}{dt} - \frac{1}{T_L} x^{(2)} \frac{dx^{(1)}}{dt} - \frac{6\beta_3}{\alpha} x^{(1)} x^{(2)} - \frac{\beta_3}{\alpha T_L} [x^{(1)}]^3 \quad (S10)$$

where

$$\gamma \equiv \frac{\beta_1 + 3\beta_3 T_L^2}{\alpha T_L} \quad (S11)$$

represents the equivalent (linearized) cooling rate, taking into account both cooling mechanisms.

For the two-laser illumination considered here, the optical intensity absorbed in the graphene is given by

$$I(t) = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \Omega t \quad (S12)$$

where I_1 is the absorbed intensity of laser 1, I_2 is the absorbed intensity of laser 2, and $\Omega \equiv \omega_1 - \omega_2$ is the heterodyne beat frequency between the two lasers.

Substituting this expression into Eq. (S8), one can find a solution for $x^{(1)}(t)$, which is used in turn to find $x^{(2)}(t)$ from Eq. (S9), and $x^{(3)}(t)$ from Eq. (S10).

The photovoltage produced through the Seebeck effect can be expressed as

$$V(t) = rT(T - T_L) = rx(x + T_L) \quad (\text{S13})$$

where rT is the Seebeck coefficient of graphene. Substituting $x = x^{(1)} + x^{(2)} + x^{(3)} + \dots$ into Eq. (S13), evaluating only the dc component of $V(t)$, and retaining only terms up to the third order in I , one finds, after simplification:

$$V(I_1, I_2) = r \left\{ \frac{I_1 + I_2}{\alpha\gamma} + \beta_1 \frac{(I_1 + I_2)^2}{(\alpha\gamma T_L)^3} - (3\beta_3^2 T_L^4 + 7T_L^2 \beta_1 \beta_3) \frac{(I_1 + I_2)^3}{T_L^6 \alpha^5 \gamma^5} \dots \right. \quad (\text{S14})$$

$$+ 2I_1 I_2 \left[\frac{\beta_1}{(\alpha\gamma T_L)^3} - (9T_L^4 \beta_3^2 + 15T_L^2 \beta_1 \beta_3 + 2\beta_1^2) \frac{(I_1 + I_2)}{T_L^6 \alpha^5 \gamma^5} \right] \frac{\gamma^2}{\Omega^2 + \gamma^2} \dots \quad (\text{S15})$$

$$\left. - 2I_1 I_2 \left[(6T_L^2 \beta_1 \beta_3 - 2\beta_1^2) \frac{(I_1 + I_2)}{T_L^6 \alpha^5 \gamma^5} \right] \left(\frac{\gamma^2}{\Omega^2 + \gamma^2} \right)^2 \right\} \quad (\text{S16})$$

For the room-temperature conditions reported here ($T_L = 300$ K), we may make the additional approximation that $\beta_1 \ll \beta_3 T_L^2$. In this regime, the linearized cooling rate (γ) is determined primarily by supercollision cooling, even though both cooling processes contribute to the measured nonlinearity in the response. With this assumption, Eqs. (S14)-(S16) simplify to:

$$V(I_1, I_2) = r \left\{ \frac{I_1 + I_2}{\alpha\gamma} + \beta_1 \frac{(I_1 + I_2)^2}{(\alpha\gamma T_L)^3} - 3\beta_3^2 \frac{(I_1 + I_2)^3}{T_L^2 \alpha^5 \gamma^5} \dots \right. \quad (\text{S17})$$

$$\left. + 2I_1 I_2 \left[\frac{\beta_1}{(\alpha\gamma T_L)^3} - 9\beta_3^2 \frac{(I_1 + I_2)}{T_L^2 \alpha^5 \gamma^5} \right] \frac{\gamma^2}{\Omega^2 + \gamma^2} \right\} \quad (\text{S18})$$

The photoinduced voltage can be rewritten as

$$V(I_1, I_2) = a_1(I_1 + I_2) + a_2(I_1^2 + I_2^2) - a_3(I_1^3 + I_2^3) \dots \quad (\text{S19})$$

$$+ 2a_2 I_1 I_2 \left(1 + \frac{\gamma^2}{\Omega^2 + \gamma^2} \right) - 3a_3 I_1 I_2 (I_1 + I_2) \left(1 + \frac{2\gamma^2}{\Omega^2 + \gamma^2} \right) \quad (\text{S20})$$

where the coefficients a_1 , a_2 and a_3 are given by

$$a_1 \equiv \frac{r}{\alpha\gamma}, \quad a_2 \equiv \frac{\beta_1}{(\alpha\gamma T_L)^3}, \quad a_3 \equiv \frac{3\beta_3^2}{T_L^2 \alpha^5 \gamma^5} \quad (\text{S21})$$

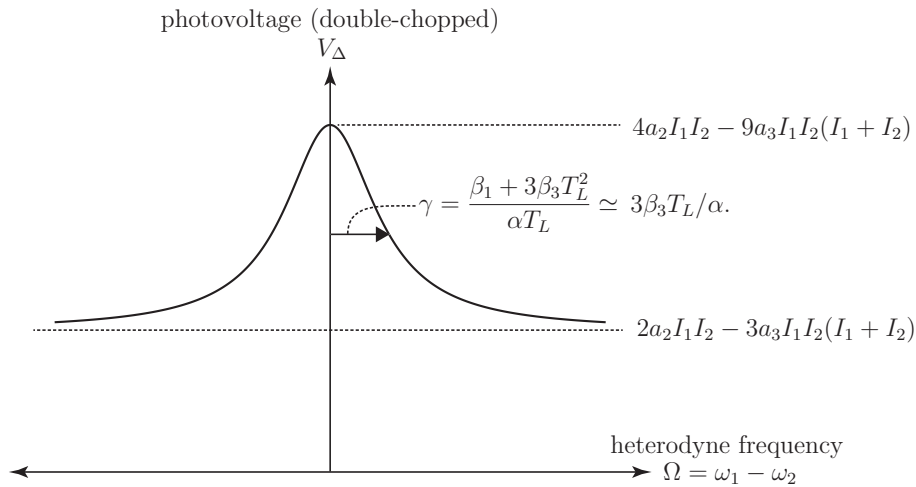


FIG. S1. dc photovoltage V_Δ as a function of the heterodyne difference frequency $\Omega = \omega_1 - \omega_2$.

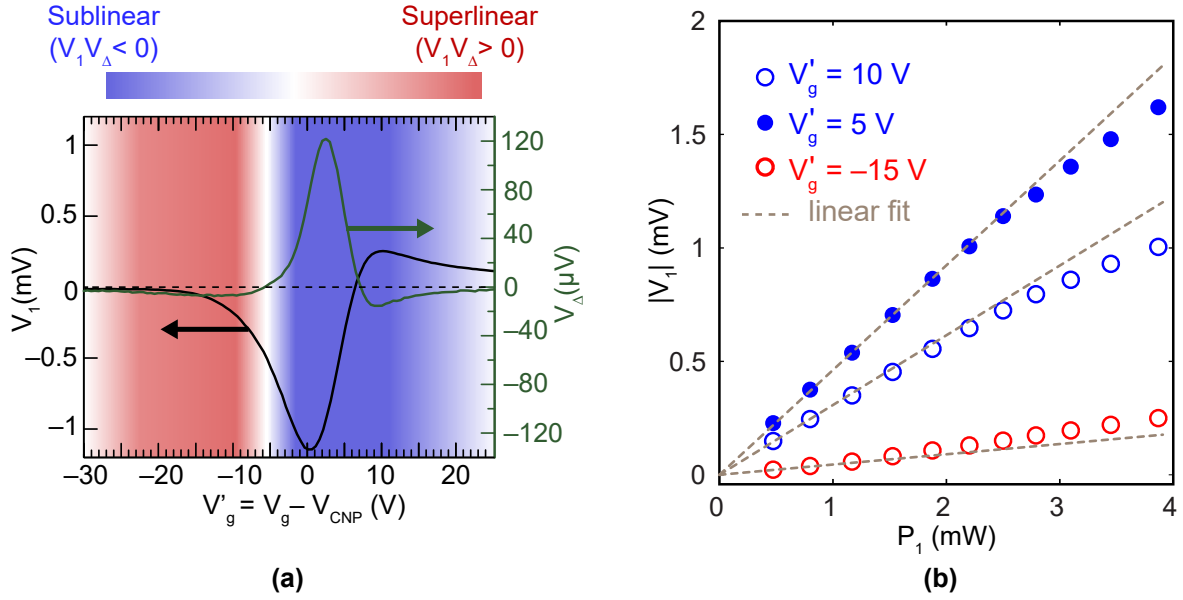


FIG. S2. (a) Single-laser photovoltage (V_1), and nonlinear photomixing signal (V_Δ) measured versus the gate voltage V_g , for the exfoliated graphene on SiO_2 . (b) Measured photovoltage $|V_1|$ versus optical input power, showing clearly the sublinear and superlinear behaviors.

When the two beams I_1 and I_2 are double-chopped and synchronously detected at the chopper difference frequency, the lock-in amplifier produces a signal proportional to Eq. (S20):

$$V_\Delta = V(I_1, I_2) - V(I_1, 0) - V(0, I_2) + V(0, 0) \quad (\text{S22})$$

$$= 2a_2 I_1 I_2 \left[1 + \frac{\gamma^2}{\Omega^2 + \gamma^2} \right] - 3a_3 I_1 I_2 (I_1 + I_2) \left[1 + \frac{2\gamma^2}{\Omega^2 + \gamma^2} \right] \quad (\text{S23})$$

The dc photovoltage therefore has a Lorentzian dependence on the heterodyne difference frequency $\Omega \equiv \omega_1 - \omega_2$, with a spectral width that is proportional to the carrier cooling rate γ , as shown schematically in Fig. S1

S2. NONLINEAR PHOTORESPONSE OF THE LOWER MOBILITY SAMPLE

Fig. S2 shows a measurement similar to Fig. 2 performed on lower-mobility exfoliated graphene on SiO_2 . In this device, the diffusion length is estimated to be only 500 nm, which is about one order of magnitude smaller than for the encapsulated device. Because of this difference, the majority of the photoresponse in this device originates from the Fermi-level pinned region near the contact, where the carrier concentration is not as easily controlled by the applied gate voltage. For positive gate voltages, Fermi level pinning produces a pn junction and charge-neutral region near the contact[34, 35], which contributes to the observed sublinear response. Otherwise, the response is qualitatively similar to that of the HBN-encapsulated device, and we observe a similar expected transition from supercollision cooling to conventional cooling under negative gate bias.

From the data in Fig. S2, the sublinear-superlinear transition occurs at $V_g = -6$ V where the estimated Fermi level is $E_F = 80$ meV. Assuming a disorder mean-free path of $l = 40$ nm (which was independently determined from dc electrical measurements), we can use equation (5) in the main text to determine the ratio of the two rate coefficients, $\beta_1/\beta_3 = 5300 \text{ K}^2$.

S3. LINEARIZED COOLING RATE AT THE CHARGE NEUTRAL POINT

The thermal model presented in the manuscript and in Eq. (S1) ignores the fact that when the graphene is gated at the charge neutral point, the carriers are no longer degenerate, and under these conditions, the specific heat (αT)

and conventional cooling coefficient (β_1) must be modified to [14, 22, 36]:

$$\alpha T \rightarrow \alpha_2 T^2, \quad \text{where} \quad \alpha_2 \equiv 18\zeta(3)k_B^3/\pi\hbar^2 V_F^2 \quad (\text{S24})$$

$$\beta_1 \rightarrow \beta_5 T^4, \quad \text{where} \quad \beta_5 \equiv 7\pi^3 k_B^5 V_D^2 / 30\rho\hbar^5 v_F^6 \quad (\text{S25})$$

and the nonlinear thermal equation under these conditions becomes

$$\alpha_2 T^2 \frac{dT}{dt} + \beta_5 T^4 (T - T_L) + \beta_3 (T^3 - T_L^3) = I(t) \quad (\text{S26})$$

If Eq. (S26) is linearized about the lattice temperature, one obtains, analogous to Eq. (S5)

$$\alpha_2 T_L^2 \frac{dx}{dt} + (\beta_5 T_L^4 + \beta_3 T_L^2) x = I(t) \quad (\text{S27})$$

where $x = T - T_L$ is the photoinduced change in electron temperature relative to the lattice. The linearized cooling rate is then

$$\gamma' = \frac{\beta_5 T_L^2 + 3\beta_3}{\alpha_2} \quad (\text{S28})$$

which is shown by the red curve in Fig. 5b.

We expect that at low temperatures, $k_B T$ will be much smaller than E_F^* , the charge-puddle-limited Fermi level, in which case the cooling can instead be accurately described by Eq. (S1). The boundary between the two regimes can be estimated by equating Eq. (S11) and Eq. (S28), which, for the parameters considered in Fig. 5 indicates that Eq. (S11) should only be applicable for $T_L < 80$ K. This condition is represented by the intersection between the blue and red curves in Fig. 5.

When the parameters determined from the low-temperature fit to Eq. (S11) are used in Eq. (S28), with no additional free parameters, we correctly predict the observed cooling rate above 80 K, which further supports the model.

S4. DEVICE FABRICATION AND DC ELECTRICAL CHARACTERIZATION

Both devices considered here employed a doped silicon substrate ($\rho_{\text{Si}} = 100 \Omega\cdot\text{cm}$), with 300 nm of thermally grown SiO_2 as a gate dielectric. The substrate served as a large-area gate contact for adjusting the carrier concentration.

The HBN-encapsulated device [20] was fabricated per the method described in [21]. A piece of polypropylene carbonate (PPC) coated polydimethylsiloxane (PDMS) was first used to pick up HBN, monolayer graphene and another piece of HBN, in that order. The resulting heterostructure was then transferred to the aforementioned SiO_2 substrate, where electron beam lithography (EBL) was used to define a hydrogen silsesquioxane (HSQ) hard mask on poly(methyl methacrylate) (PMMA). The surrounding areas were then etched in CHF_3 plasma to shape the device channel and expose the graphene edge. Afterwards, HSQ was lifted off and EBL was used again to define the contact leads and pads using PMMA, and 1.5 nm/20 nm/50 nm Cr/Pd/Au was e-beam evaporated and lifted off for edge contact. The HBN-encapsulated graphene channel length was 7 μm and width 0.7 μm .

For the second device, a single layer of graphene was mechanically exfoliated from bulk graphite and transferred directly to the SiO_2/Si substrate. The exfoliated graphene exhibits a mobility about $\mu = 5,000 \text{ cm}^2 \text{V}^{-1} \text{s}^{-1}$, which was inferred from dc transport measurements. Electron-beam lithography was used to pattern a bi-layer resist comprised of methyl methacrylate (MMA) and polymethyl methacrylate (PMMA). The contacts were deposited using successive angled evaporations of chromium (15 nm) and gold (30 nm), thereby providing dissimilar contacts to the opposing edges of the graphene channel. Dissimilar electrical contacts are not necessary when the optical beams are focused onto one contact, as for the measurements reported here, but this configuration also provides the thermal asymmetry needed for detection of spatially homogeneous or longer wavelength illumination. The graphene channel length was 2.5 μm and width 7 μm .

To quantify the electrical characteristics and gating behavior, we conducted unilluminated measurements of the dc resistance as a function of the gate voltage, for both the HBN-encapsulated device and the non-encapsulated device. Fig. S3 shows the dc measurements, along with optical micrographs showing the graphene active region, contact geometry, and cross-sectional diagram.

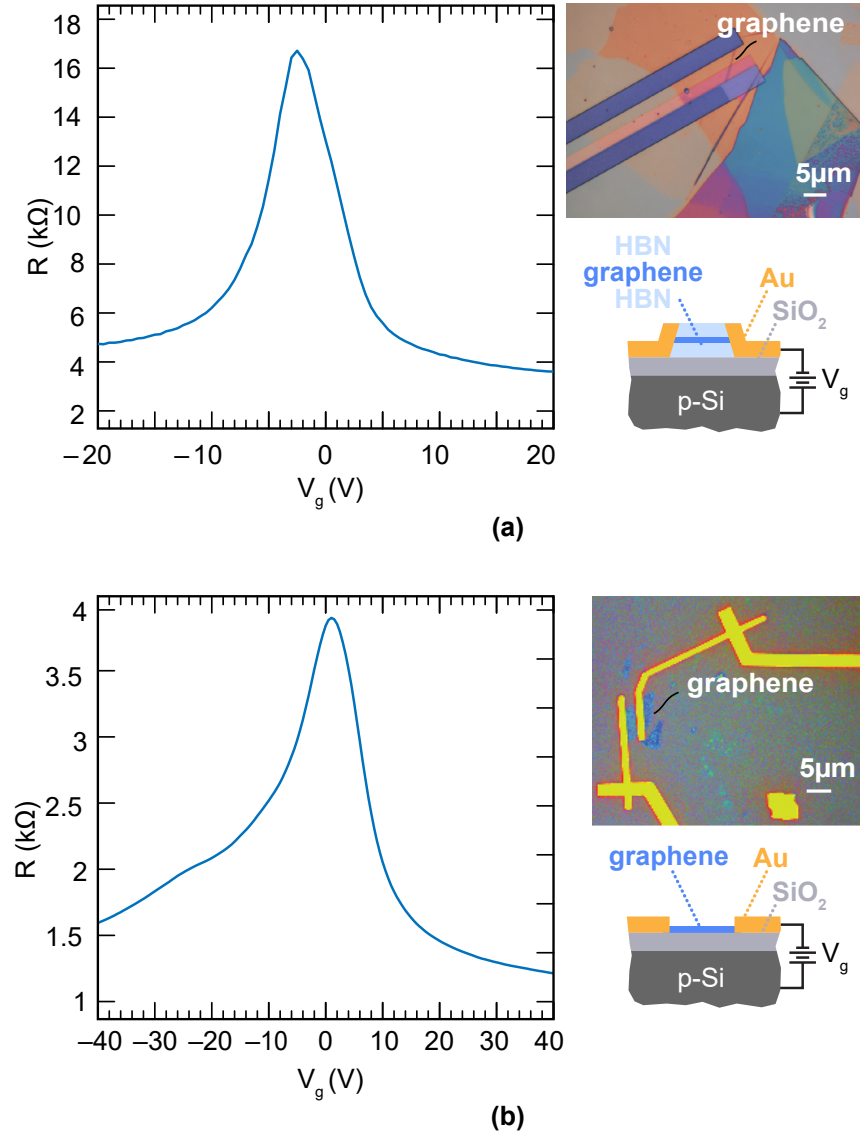


FIG. S3. The dc resistance R as a function of the applied gate voltage V_g and the optical micrograph for (a) the HBN-encapsulated graphene device and (b) the exfoliated graphene on SiO_2 device.